Hidden Markov Trees of Wavelet Coefficients for Transient Detection in Audiophonic Signals

Laboratoire d'Analyse Topologie et Probabilités / CMI
Université de Provence
39, rue F. Joliot Curie, 13453 Marseille Cedex 13, France

http://www.cmi.univ-mrs.fr/~molla/GTS/index.htm

Stéphane Molla & Bruno Torrésani *

Abstract

We describe in this paper a new way to model and estimate transient features in audiophonic signals. The approach is based upon Hidden Markov Models, following a dyadic orthogonal wavelet decomposition. The distribution of coefficients depends on a hidden state ("transient type" or "residual type" coefficients) which is estimated in order to proceed to a selection. Details about this modeling are given, and some results on transient extraction in the context of an hybrid audio coding scheme are discussed.

keywords: wavelet analysis, Markov models, binary tree, audio transient, scale law

1 Introduction

Hybrid audiophonic signals models have received an increasing attention during these last years. Recent developments appeared for signal modeling as well as for coding and compression aims, based on the following idea: a sharper modeling of the signal may lead to substantial gains in the coding rate. In this framework, we take the model:

$$signal = tonals + transients + residual$$

as proposed in [1, 2], where an audio signal is supposed to have three additive "independent" parts. Each component is then estimated and encoded separately by adaptive transform coding. In such an approach, a component is characterized by the most significant coefficients given by a well-chosen basis decomposition: a local trigonometric basis for tonals, and a wavelet basis for transients, in order to estimate each part as accurately as possible. Eventually, we aim to model the final residual signal as a stationary random process.

Our strategy is first to estimate a tonal part, encode it and substract it from the original one. This paper deals with the estimation of a transient part in this residual signal. We use an orthonormal wavelet expansion of this signal in order to select the "transient-type" coefficients, namely coefficients following the property (cf. [1]): *a transient structure is a connected tree of wavelet coefficients which satisfies a given relevance property.* This property is defined further in probabilistic terms via the model imposed for the selection, and allows us to exploit the underlying hierarchical organization of the wavelet coefficients over a dyadic tree structure (see Figure 1).

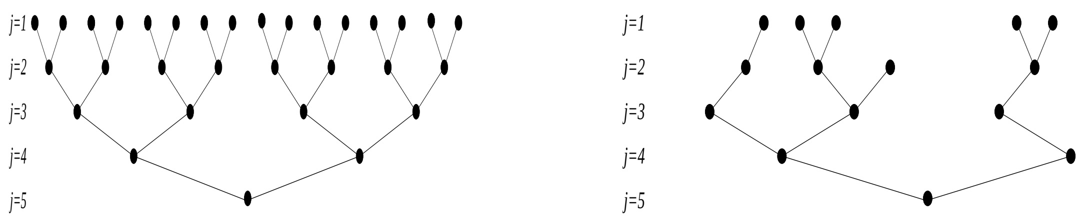

Figure 1: Dyadic grid of wavelet coefficients, and a subtree of relevant coefficients.

*{molla,Bruno.Torresani}@cmi.univ-mrs.fr

Our transient model is based upon the HMT model (Hidden Markov Tree model) proposed in [3] in the context of signal and image modeling and denoising, and adapted to our study.

2 Model

2.1 Wavelet transform

A wavelet transform is a multi-scale decomposition of a signal. Let ψ be a compactly supported wavelet, and set as usual $\psi_{jk}(t) = 2^{-j/2}\psi\left(2^{-j}t-k\right)$. Denote by $d_{jk} = \langle x, \psi_{jk}\rangle$, $j = 1, 2, \dots J$ the wavelet coefficients of a signal $x \in L^2(\mathbb{R})$. The latter are obtained via a sub-band coding algorithm with complexity $\mathcal{O}(N)$ for length-N signals (see [5] for details), and are naturally associated with a dyadic tree structure: each coefficient d_{jk} at scale j has two children $d_{j-1,2k}$ and $d_{j-1,2k+1}$ at scale $j-1$.

According to the common practice, the samples x_k of the input signal are identified with small scale scaling function coefficients, and we consider wavelet expansions of the form

$$\langle x, \phi_{0k}\rangle \approx x_k \ , \quad x = \sum_k s_k \phi_{Jk} + \sum_{j=1}^{J}\sum_k d_{jk}\psi_{jk} \ .$$

which naturally leads to a dyadic tree structure for coefficients d_{jk}.

As a transient feature is well-localized both in time and frequency, we will choose a wavelet according to this characteristic. Anymore, we will take as a definition that:

Definition 1 (transient feature) *a transient feature is a* **connected** *wavelet coefficient tree whose root corresponds to the largest considered scale of the decomposition.*

In order to better understand this definition and the choice of a wavelet basis, let us emphasize some key properties for wavelet expansions (and coefficients) which are shared by most of transient structures in audiophonic signals (recall that each atom ψ_{jk} of the wavelet expansion is localized both in time and frequency):

- if a coefficient is large (resp. small), its neighbors are likely to be large (resp. small),
- large values (resp. small) tend to propagate through scales.

2.2 Adapted HMT model

Following the so-called HMT model in [3], a random variable $D_{j,k}$ (representing a wavelet coefficient) is associated to each node (j,k) (j is the scale, k the position) of the resulting dyadic tree of a wavelet expansion (cf. section 2.1). Here are the details of the model:

- variables $D_{j,k}$ follow a distribution depending on a hidden state $X_{jk} \in \{T, R\}$ (where "T" stands for Transient, and "R" for Residual),
- in each scale j, coefficients $D_{j,k}$ hare normally distributed, with zero mean and a variance conditioned by the associated hidden state:
  - T-type coefficients follow a distribution with "large" variance $\sigma^2_{T,j}$,
  - and R-type ones a "small" variance $\sigma^2_{R,j}$.
- hidden states distribution is also a random process given by a "coarse to fine" Markov Chain on the tree with 2×2 transition matrix Π. A further constraint is imposed:

  $$\boxed{R \to T \textit{ transition is } \textbf{forbidden}}$$

  (which means: $\mathbb{P}\{X_{child} = T | X_{parent} = R\} = 0$).

  This condition ensures the connectedness of transient coefficient trees, and therefore directly yields transient structures in the sense of Definition 1. Moreover, the transition matrix Π is fully defined by the *scale perstistence probability* $\mathbb{P}\{T \to T\}$:

  $$\pi = \mathbb{P}\{T \to T\} = \mathbb{P}\{X_{child} = T | X_{parent} = T\}$$

  The transition matrix then assumes the form:

  $$\Pi = \begin{pmatrix} \pi & 1-\pi \\ 0 & 1 \end{pmatrix}$$

Finally, if we take an initial probability distribution $\nu = (\nu_T, \nu_R)$ at coarsest scale J, our model is fully characterized by the set of parameters:

$$\Theta = \{\nu,\, \pi,\, \sigma_{T,j},\, \sigma_{R,j},\; j \;=\; 1 \ldots J\}\ ,$$

and we can give a probabilistic definition of a *transient feature* as:

Definition 2 (probabilistic transient feature) *a transient feature is a connected tree of* T**-type** *wavelet coefficients.*

With such a characterization, we can say that a transient feature is a signal which has a wavelet expansion following a process with the same law at each scale (a normal law with zero mean) with different variance by scale (the underlying T-type process, with parameters which can be viewed as a signature of the transient).

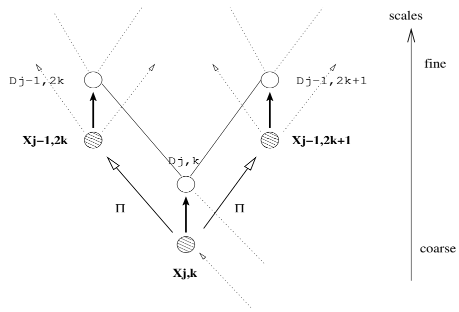

Figure 2: Graphical representation of the Hidden Markov Tree model

In other words, given a realization of this process, the corresponding "transient" part of the so-generated signal is the wavelet synthesis from T-type coefficients, according to Definition 2:

$$x_{tr} = \sum_k s_k \phi_{Jk} + \sum_{j,k;X_{jk}=T} d_{jk}\psi_{jk}$$

Remark: Notice that we have included the low-pass component s_k in the transient reconstruction, even though it cannot be really termed "transient". However, this coefficient does not have a large cost in terms of coding and is rather weak for audio signals.

With applications to audio signal compression in mind, it is interesting to investigate basic quantities such as the average number of T-type coefficients, or the corresponding energy. For instance, the parameters Θ of the hiddden states distribution control the so-called rate, *i.e.* the number of bits needed to encode the transient signal.

Proposition 1 (rate control) *Let a set of parameters* Θ *be given, and let* $(N_j)_{j=0\ldots}$ *be the random process giving the number of* T*-type coefficients at scale* j *(assuming that* $j = 0$ *corresponds to the root of the tree). Let* N *be the total number of* T*-type coefficients in a complete tree of scale depth* J*. Then we have:*

$$\mathbb{E}\{N_j\} = \nu(2\pi)^j\ ,$$

and

$$\mathbb{E}\{N\} = \begin{cases} \nu \times \frac{1-(2\pi)^J}{1-2\pi} & if \pi \neq 0.5, \\ \nu J & if \pi = 0.5. \end{cases}$$

proof: The random variables N_j follows a Galton-Watson branching process with parameters π and ν. And we have $N = \sum_{j=0}^{J-1} N_j$, giving $\mathbb{E}\{N\} = \sum_{j=0}^{J-1} \mathbb{E}\{N_j\}$. We refer to [7] for an example. $\square$

Following Prop. 1, we can now have an estimation of the extracted transient signal energy given by the following result (whose proof follows the same lines as above)

Proposition 2 (energy control) *With the same notations as before the average energy of a transient feature reads*

$$\mathbb{E}\left\{ \sum_{j,k;X_{jk}=T} |d_{jk}|^2 \right\} = \nu \sum_{j=0}^{j=J-1} \sigma_j^2 (2\pi)^j$$

More results and more general norm estimates will be discussed in a forthcoming publication, together with a comparison with the deterministic transient estimation algorithm given in [2].

Still keeping in mind the problem of encoding the transient component of the signal, let us remind that the coefficients have to be encoded, as well as their addresses (the so-called *significance map*). The significance map encoding is another good argument in favor of tree-based approximations. The following proposition gives au upper bound for the rate of significance map encoding.

Proposition 3 (tree encoding) *With the same notations as before, let R_{SM} denote the number of bits necessary to encode the significance map of a transient wavelet coefficients tree. Then we have*

$$\mathbb{E}\{R_B\} \leq \begin{cases} 2\nu \times \frac{1-(2\pi)^J}{1-2\pi} & if \pi \neq 0.5, \\ 2\nu J & if \pi = 0.5. \end{cases}$$

proof: We exploit the connected tree structure of the coefficients. If a coefficient is not retained, all its children are not retained, then we only order the tree from the root and use one bit per retained coefficient, and one bit for each root of a non-retained coefficients tree (compared with zerotrees introduced by Shapiro [8]). This shows that we need at most 2 bits per coefficient to encode retained coefficients adresses. The proposition then directly follows from Proposition 1. □

2.3 Transient estimation

According to our model, we have to consider the two classical estimation problems:

- parameters estimation: estimate the best set of parameters Θ in the maximum likelihood sense, *i.e.* maximizing $\mathcal{L}(\theta) = \log \mathbb{P}\{D|\Theta\}$,
- states estimation: find the optimal distribution of hidden states according to our best set of parameters.

The first problem is solved with a classical EM (expectation - maximization) algorithm, which allows us to compute the *a posteriori* probabilities values $\mathbb{P}\{X_i = S\,|\mathcal{T}_1\}$ and $\mathbb{P}\{X_i = s_1,\, X_{\rho(i)} = s_2|\, \mathcal{T}_1\}$ where $S, s_1, s_2 = T, R$, for a full given tree $\mathcal{T}_1$ of wavelet coefficients. States estimation can be computed with an adapted Viterbi algorithm (see [6] for an introduction and [4] for further details). Note that hidden states estimation directly yields estimates for the transient features.

2.3.1 EM algorithm

As seen before, parameters estimation leads us to use a EM procedure to solve this problem. Such an algorithm is well-known in the context of HMCs, based on the computation of the so-called “forward” and “backward” variables. Such a scheme has been developped in [3] for Hidden Markov Trees (“upward-downward” algorithm) but, as in the case of HMCs, is limited by numerical underflow problems due to the underlying multiplicative procedure[1]. This algorithm has a computational complexity $\mathcal{O}(nK^2)$ where n is the size of the tree (number of nodes) and K the number of hidden states (2 in our case).

A solution to these limitations, based on the same model, has been proposed by Durand and Gonçalvès in [4]. It consists of an algorithm inspired by Devidjer's “conditional forward-backward” recursion (1985), with the same complexity as the first one but more robust to underflows.

Moreover, it leads to a MAP (Maximum A Posteriori) algorithm for this model, giving a solution to the second problem of hidden states restoration, which is discussed in next section.

2.3.2 Estimation procedure

The natural procedure used in HMC strategies to find the optimal distribution of hidden states is the Viterbi algorithm. It yields the best sequence of hidden states for a given set of parameters.

Transposing the same strategy in our case would be too computationally expensive because we would have to consider permutations of states at each scales of our decomposition, with a set of nodes exponentially increasing with the depth of the tree. Fortunately, a MAP algorithm has been developped in [4] for this model, with graphical arguments involved in the proof. This algorithm maximizes $\mathbb{P}\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\,|\mathcal{T}_1, \Theta\}$, giving the most likely hidden states repartition given a realization $\mathcal{T}_1$ and the model Θ, with a computational complexity order of $\mathcal{O}(nK^2)$.

[1] the numerical problems manifest themselves even if we operate to a rescaling, *i.e.* if we try to give a common numerical scale to each variable.

The Viterbi algorithm yields the most likely distribution of hidden states according to the model. Nevertheless, this scheme is based on the assumption of independence between tonal and transient parts, which is reasonable but not physically true. Indeed, audiophonic signals are very complex process and our modeling is expected to be as general as possible. That's why, as we will see in the discussion below (section 3), even though the Viterbi algorithm yields very impressive results, an alternative may also considered in order to obtain relevant results into terms of compression[2], rate control[3] and psychoacoustic features[4].

In order to obtain this control, hidden states estimation may be performed through *a posteriori* probabilities thresholding, *i.e.* only wavelet coefficients i such that $\mathbb{P}\{X_i = T\,|\,\mathcal{T}_1, \Theta\} > p_0$ are retained. The aim of this thresholding is to select only coefficients i belonging to state T, *a.s.*, according to our model. In our simulations, we have used the value p_0 as a subjective choice motivated by the consideration of the limited numerical precision of computers and the expected amount of coefficients to be retained. Indeed, in such a multiplicative algorithm, we cannot expect to select precisely coefficients such that $\mathbb{P}\{X_i = T\,|\,\mathcal{T}_1, \Theta\} = 1$, which is the expected procedure. In a signal coding context, the threshold probability p_0 may easily be adapted to any prescribed bit rate to be allocated to the transient component.

2.3.3 Application to transient detection

A natural extension of this model in the context of audiophonic signal processing is the re-estimation of parameters Θ over a set of several trees, say N. Indeed, a transient feature is defined in contrast with a "general behavior" of the signal, and not only in the context of a single window (by frame or window, we refer to a set of samples used to compute a full wavelet expansion leading to a tree of coefficients with a fixed given depth).

To obtain these trees of coefficients, we compute a wavelet decomposition for $N \times 1024$ samples of the signal, over 10 scales depth, what give our N full trees. Note that in practice we make the expansion over $(N + 2n) \times 1024$ samples, where n depends on the filter length associated to our decomposition. Then we only consider the N center frames to avoid edge effects problems.

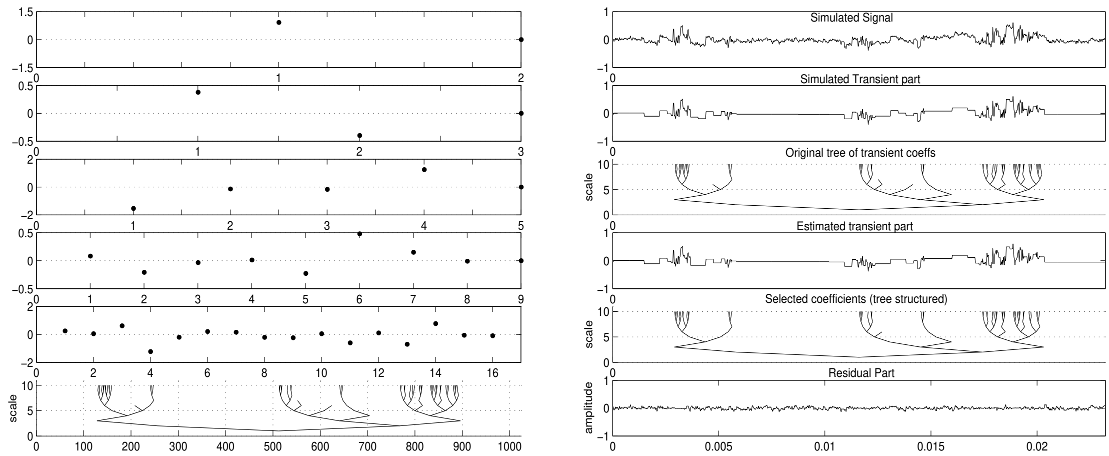

Figure 3: Left: Simulated tree of wavelet coefficients according to our model (147 coeff., $\pi = 0.75$). From top to bottom: generated coefficients from coarsest to finer scales (first five scales), and full tree. Right: various plots (139 coeff. estimated, $\pi_{est.} = 0.781$)

Thus, in such a strategy, the set of parameters Θ will be the same within N consecutive trees, what leads to the problem of how to estimate this set in a general way, and how (and when) to change this set of parameters?

The answer to the second question only depends on the choice of N, and we will focus here on the first one. We present here two different methods for Θ estimation over N frames:

[2]remember that this extraction is a part of a global scheme of signal compression, what imposes a constraint over the number of retained coefficients.

[3]an efficient and parametric coder will have a way to control the output rate of the encoded signal.

[4]selected transients must be perceptually close to the original one (transparent coding), and inaudible structures don't need to be selected and encoded.

- a global method: Θ parameters are re-estimated simultaneously during the EM loops over the whole set of N frames via a procedure of "tree tying", *i.e.* using EM-variables and *a posteriori* probabilities computed within each tree,
- a local method: best Θ set is computed in each frame (*i.e.* for each tree) and a global set is estimated given these sets, e.g. via the formula $\hat{\sigma}_{S,j} = (\sum_{k=1}^{N} \hat{\sigma}_{S,j}^{p}(k)/N)^{1/p}$ for variances.

2.3.4 Simulation results

We first present results obtained on synthetic signals (cf. Figure 3), *i.e.* a synthesized tree of wavelet coefficients with a fixed given set of HMT parameters. We clearly obtain in this case very good estimates as proved by a comparison beetween simulated and estimated trees. We have generated here a 10 scales depth tree corresponding to a wavelet decomposition of a 1024 samples long signal, which is the typical length we use for real-world audio signals sampled at 44.1 kHz.

Figure 4 gives some results for the global procedure of estimation and selection over a synthetic signal with a given set of parameters. In this example, 12 windows of 1024 samples are considered, what means a wavelet expansion over 10 scales depth. Again, we have very good results in the estimation of transient features as shown by a comparison between simulated and estimated trees. Notice that the residual part (original signal with removed estimated transient part) is rather stationary.

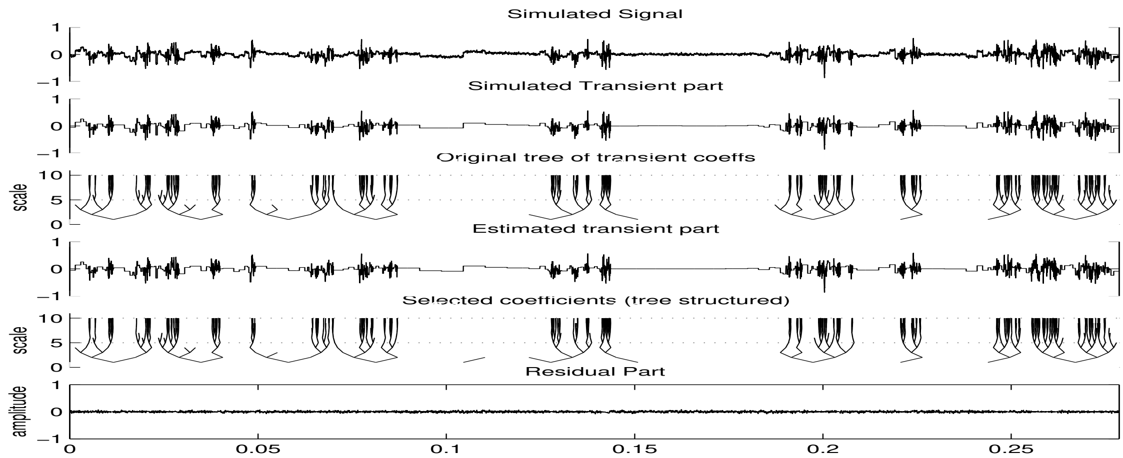

Figure 4: From top to bottom: Simulated signal, transient, trees and and estimated transient and trees via the global procedure of estimation, residual part (signal with estimated part removed) // simulated: 1331 coefficients (*i.e.* 10.8%) with $\pi = 0.75$, estimated: 1265 coefficients, $\pi_{est.} = 0.743$

3 Application to audio signals and discussion

In this section, we present some results for transient estimation and extraction on audio signal. We use here a jazz signal (containing drums, (double) bass, brass, ...) sampled at 44.1 kHz (audio CD quality) of about 2.5 s length from which a tonal part has been previously removed by an appropriate coefficients thresholding after a local trigonometric decomposition as introduced in section 1. Thus we consider our signal to be fairly "non-tonal" and compute our two estimation procedures with $N = 10$ frames of width 1024 samples (*i.e.* max depth $J = 10$ per frames of about 23 milliseconds long) for the wavelet decomposition (and so for trees).

In Figure 5, we estimate parameters with the global procedure (tree tying) before to estimate the transient part via a thresholding of *a posteriori* probabilities, as presented in section 2.3.2.

In this case (with threshold probability $p_0 = 0.99$), 19% of coefficients are retained. A good estimation is obtained in sound quality terms. We can see that after extraction of the transient part, the residual part is fairly stationary, at least within time frames of reasonable length, which is a good point for a further modeling.

In Figure 6, we again estimate the parameters with the global procedure before estimating the transient part via the Viterbi algorithm. We thus obtain estimated T-type coefficients in the Maximum Likelihood sense, what corresponds to 27% of the wavelet coefficients. The major difference with previous example is the stronger estimation of each transient structure. This is reflected by a

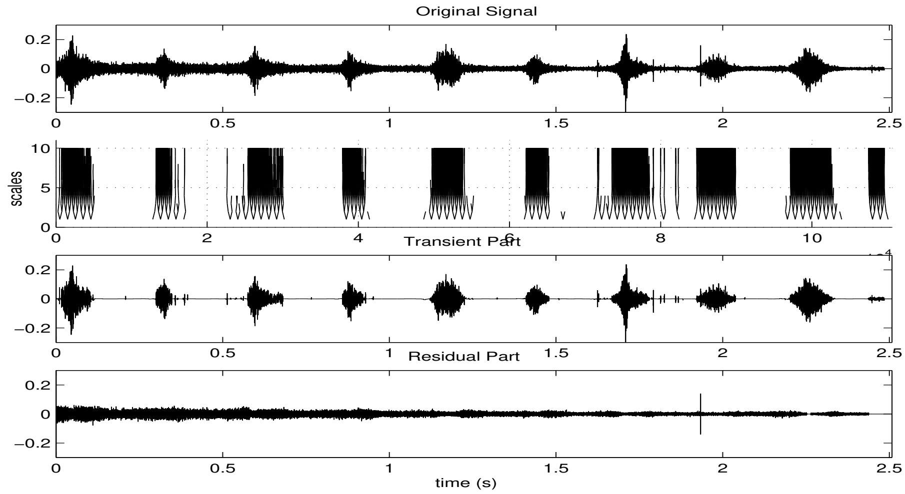

Figure 5: Transient estimation/extraction: global parameters re-estimation, *a posteriori* probability thresholding ($p_0 = 0.99$).

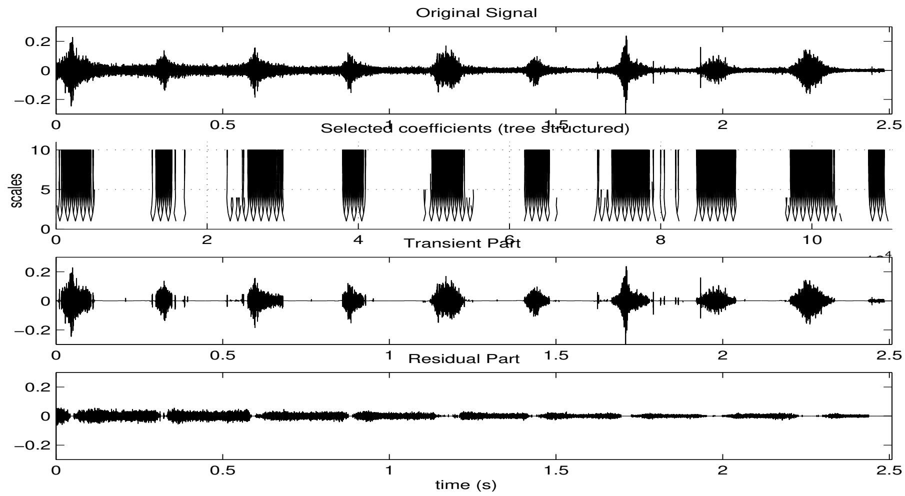

Figure 6: Transient estimation/extraction: global parameters re-estimation, Viterbi state estimation.

stronger persistence of tree branches through the scales, and therefore leads to a larger number of selected coefficients. Therefore, a phenomenon not present in Figure 5 is to be noted here: the presence of "holes" in the residual part, that are consequences of quasi-full trees of selected coefficients (overall in the most powerful part of the transient features). However, the Viterbi estimation is much more precise than a thresholding, as we can see in Figure 7, because of its ability to capture small and very localized structures.

We do not provide here illustrations or further details on the so-called local method, because most relevant performances in terms of transient estimation are obtained with the global method. Nevertheless, this method has some interesting advantages. On one hand, in a coding scheme perspective, better control over transient bit rate can be performed with this method with appropriate parameters of re-estimation. This control is very helpful to have precise and predictable measures of the behaviour of the coder. On the other hand, this method performs better in terms of computational burden because of its faster convergence during the parameters estimation phase. Indeed, this algorithm does not need an overall convergence as in the case of tree tying involved in the global method. This point can also play an important role in the choice of the method even if the key point is the reconstruction time of our signal.

Remark: It is important to stress at this point that high computational cost is not necessarily a major drawback for audio encoders. The most important point is that the decoder be compatible with real time processing. A number of applications are compatible with computationally intensive encoding but efficient decoding, and this is precisely the case of the coding scheme we have been

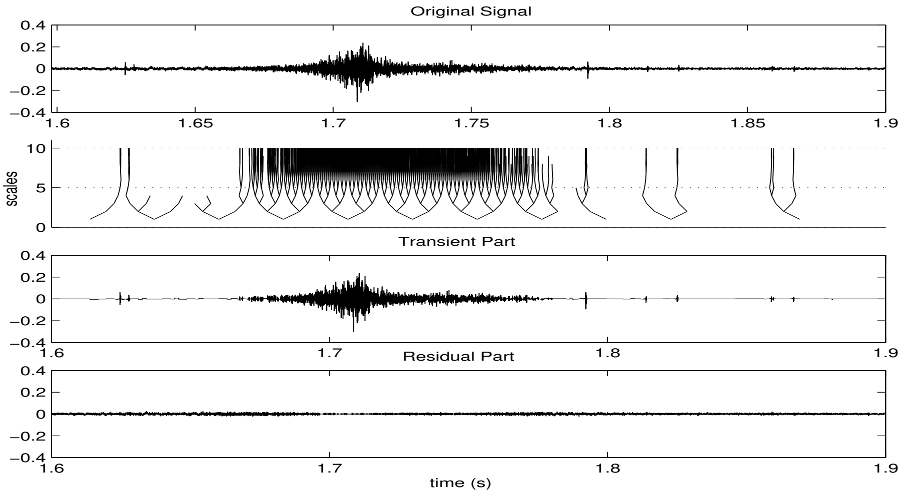

Figure 7: Detail of Figure 6 around time 1.75s.

outlining in this paper.

4 Conclusion

The combination of wavelets with a probabilistic model for transient modeling and estimation in the context of audiophonic signals gives interesting results, in terms of both transient modeling and coding rate control. On one hand, this model seems rather realistic, at least for "attacks" and others important transients in sounds. Thus, explicit use of these structures improves the quality of the estimation. On another hand, in the context of coding, considerable savings are obtained by structuring datas. Indeed, as we are in a non-linear transform coding scheme (*i.e.* retained coefficients are estimated and not known *a priori* as in some codec), tree structure allows us to reduce drastically coefficients address encoding.

Finally, the general character of this model combined with the ability of estimation of the parameters allows us to cover a wide range of audio signals with significant results.

Now, to obtain a complete and self-consistent codec based upon the same ideas, the development of an efficient algorithm of tonal estimation is under progress. The goal is to optimally couple tonal and transient extractions, whose performances are closely related. The use of high-level psychoacoustic criterions may help improving each layer estimation and encoding.

References

[1] L. Daudet, S. Molla, and B. Torrésani, "Transient modeling and encoding using trees of wavelet coefficients", in *Proc. 18th Symposium GRETSI'01 on Signal and Image Processing, Toulouse*, September 2001. available at: `http://www.cmi.univ-mrs.fr/~molla/index.htm`

[2] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding", *Signal Processing*, 2002, Special issue on Image and Video Coding Beyond Standards, (*to appear*).

[3] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based signal processing using hidden Markov models", *IEEE Transactions on Signal Processing*, 1998, Special Issue on Filter Banks.

[4] J.B. Durand and P. Gonçalvès, "Statistical inference for hidden markov tree models and application to wavelet trees", Technical Report 4248, Institut National de Recherches en Automatique et Informatique, September 2001.

[5] S. Mallat, "A wavelet tour on signal processing", *Academic Press*, 1998.

[6] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition", *Proceedings of the IEEE*, vol. 77, pp. 257–286, 1989.

[7] S. Karlin, "Initiation aux processus aléatoires", *Dunod Ed.*, 1969.

[8] J. M. Shapiro, "Embedded Image Coding Using Zerotrees of Wavelet Coefficients", *IEEE Trans. Signal Processing*, vol. 41, n.12, pp. 3463–3479, 1993.